行政院國家科學委員會專題研究計畫成果報告

使用類神經網路來作文件自動分類之研究
A Study on Automatic Document Classification by Using Neural Networks

計畫編號：NSC 86-2213-E-032-003
執行期限：85 年 8 月 1 日至 86 年 7 月 31 日

主持人：洪文斌　淡江大學資工系 (E-mail: horng@cs.tku.edu.tw)
共同主持人：黃連進　淡江大學資工系 (E-mail: micro@cs.tku.edu.tw)

一、中文摘要

本研究計畫利用類神經網路的強大學習能力，來學習文件的內在分類規則，以期達到文件自動分類的目的。在本研究中，我們採用 *ACM Computing Reviews* 的分類法作為分類的依據。我們從該期刊共收錄了 56 個中類別，6424 篇論文為實驗用資料。再以其中的論文題目和出處當作該文件的素描。取其中十分之一為測試資料，其餘為訓練資料。在實驗前，首先將所有英文單字中，名詞單複數，及動詞單複數和其現在分詞、過去式、過去分詞取其原型；並將文件出處視為一英文單字。我們從訓練資料中，共選出了 1146 個關鍵詞。在原先訓練文件中，有 26 篇沒有出現任何關鍵詞，去除後剩下 5755 篇；相同的，原先測試文件中，有 6 篇沒有出現任何關鍵詞，去之剩下 637 篇，此為以下實驗用之資料。在類神經網路的分類實驗上，我們實驗了倒傳遞網路和學習向量量化網路。其實驗結果分別為：在倒傳遞網路方面，訓練資料召回率為 70.7%，測試資料為 57.1%；在學習向量量化網路方面，訓練資料召回率為 86.9%，測試資料為 50.5%。此外，我們並實驗了傳統的機率模式、向量模式、和 Hamill 的分類結果以為比較。實驗結果顯示，類神經網路分類結果均優於上述的傳統方法。

關鍵詞：類神經網路，文件自動分類，資訊檢索

Abstract

This project uses the powerful learning capability of artificial neural networks to learn the intrinsic classification rules for categorizing documents automatically. In this research, the classification system of *ACM Computing Reviews* is based on. Totally 6424 papers, including 56 classes, are collected from it. The title and its source of each paper are used as its document profile. Among the collected papers, 10% of them are used as test data, and the remaining are used as training data. Before experiments, the stemming process is performed for all papers, including plurals, past tenses, and gerunds. There are 1146 keywords selected from the training data. Documents without any keywords are deleted. The training data and test data remain 5755 and 637 papers, respectively, used for the following experiments. Two network models, back-propagation networks (BP) and learning vector quantization networks (LVQ), are performed for document classification. The experiment results show that in BP, the recall rates of training data and test data are 70.7% and 57.1%, respectively; in LVQ, they are 86.9% and 50.5%, respectively. Finally, we also perform the experiments of traditional probability model, vector space model, and Hamill's method to obtain the results for comparison with neural network models. The experimental results indicate that neural networks for document classification are all superior to traditional methods.

Keywords: artificial neural networks, automatic document classification, information retrieval

二、緣由與目的

隨著資訊時代的來臨，網路的發達，資訊正以等比級數般的數量在激增。要在這龐大的資料中，找尋相關的資訊，確非易事。因此，文件自動分類的研究便應運而生。文件自動分類的目的即是利用電腦將性質相近的的資料或文件排放在一起，以提高文件分類的正確性與一致性，便於使用者能夠快速地檢索到相關的資訊。

Maron [10] 於 1961 年發表的論文，應該是文件自動分類領域中最早的文獻。Maron 認為，對於所要分類的文件，我們可以從文件中的某些詞找到分類的線索，稱之為關鍵詞(keywords)。若電腦也能從文件中自動找出這些關鍵詞，那麼便可以做到所謂的自動分類。在該論文中，他首先挑選了 405 篇文件，其中的 260 篇是訓練資料，另外的 145 篇是測試資料。每篇均取其摘要當作文件的素描。結果，在所有訓練資料中，共得到 3263 個不同的詞。其次，做關鍵詞篩選，把這些詞中去掉頻率最高的 55 個，及只出現一次或兩次的詞，便剩下 1088 個詞。再根據 Entropy 公式來計算，看這些詞在文件的分佈情形。只有分佈不平均的才有分類的價值，所以把平均者去掉。最後就只剩 90 個詞，也就是關鍵詞。他採用機率模式來作文件分類的實驗。結果顯示，在扣除不含關鍵詞及只含一個關鍵詞的文件後，訓練資料中有 84.6%的召回率，而測試資料中也達到 51.8%。(召回率即系統辨識正確之文件數和文件總數之比率。)

之後，陸續還有許多學者提出不同的作法，像 Borko 和 Bernick [2]延續了 Maron 的實驗，嘗試用向量模式來做分類；Kar 和 White [6]的實驗，提出了第二選擇類別來提高正確率及循序的演算法來節省時間及空間；Kwok [8]的分類實驗，除了論文題目及摘要外，他還利用論文所引用參考文獻的題目作為分類之用；以及 Hamill 和 Zamora [3]的分類實驗(以下簡稱為 Hamill 方法)，提出了只用文件的題目來做分類。近年來的研究也加入了統計分析、專家系統、和自然語言處理等先進的技術，以提高分類的正確性[1,4,5]。

類神經網路自從 1980 年代復甦以來，已吸引了許多研究者的投入，從事基礎性理論的研究與實務性的應用，並獲致相當的成功。近年來，也開始有人嘗試利用類神經網路來作文件的聚類(clustering)和分類 (classification) 之研究。MacLeod 和 Robertson [9]利用非監督式學習網路中類似自適應共振理論網路 (Adaptive Resonance Theory Networks)模式來作文件之自動聚類研究。而 Yang [13]依據傳統的相似性公式(cosine similarity measure)來設計一三層的類神經網路架構，其輸入層為文件中所出現的單字，輸出層為分類的類別，而隱藏層為訓練之文件。輸入層到隱藏層間的連結加權值(link weights)根據相似性公式適當的給定，而隱藏層到輸出層之間的連結加權值則為一條件機率 P(類別|文件)。對一測試文件相對於某一類別的相關性可輕易地由上述的相似性與條件機率相乘而得。

本研究計畫的主要目的，是嘗試利用類神經網路的強大學習能力，來學習文件的內在分類規則，以期達到文件自動分類的目的，並提高其分類的正確性。在本研究中，我們利用倒傳遞網路[12]和學習向量量化網路[7]兩種不同監督式學習網路模式來作文件自動分類之實驗。此外，我們並實驗了傳統的機率模式、向量模式、和 Hamill 的分類結果以為比較。實驗結果顯示，在訓練資料召回率方面，學習向量量化網路高達 86.9%，表現最優異；而在測試資料方面，倒傳遞網路可達 57.1%，拔得頭籌。此等顯示了類神經網路應用在文件自動分類上的優越性。

三、實驗方法及步驟

在本文件自動分類研究中，我們採用 *ACM Computing Reviews* 的分類作為分類

的依據。其分類系統共有 11 個大類和 80 個中類。我們從該期刊上，收錄了自西元 1986 年 1 月份起至 1997 年 6 月份止，共 67 個中類別，6507 篇論文。其中有 11 個中類別所含之論文數少於 10 篇，沒有足夠資訊以為訓練，將之刪除，並將有重覆出現之論文 49 篇去除，剩餘 56 個中類別，6424 篇論文為實驗用資料。再以其中的論文題目和出處當作該文件的素描。依論文收錄之順序，每第十篇取作為測試資料，計有 643 篇文件；其餘有 5781 篇為訓練資料。

在實驗前，首先將所有英文單字中，名詞單複數，及動詞單複數和其現在分詞、過去式、過去分詞取其原型；並將文件出處視為一英文單字。在關鍵詞的選取中，我們從訓練資料中，先去除 235 個 stop words (如 about, but, in, not, ...) 及只包含一個字元的單字，再以單字至少出現 5 次以上和其 Entropy 值小於等於 $\log_{10}(20)$ 為標準，共選出了 1146 個關鍵詞。在原先訓練文件中，有 26 篇沒有出現任何關鍵詞，去除後剩下 5755 篇；相同的，原先測試文件中，有 6 篇沒有出現任何關鍵詞，去之剩下 637 篇，此為以下實驗用之資料。

在類神經網路的文件分類實驗上，我們利用 NeuralWare 公司的 NeuralWorks 類神經網路軟體[11]，實驗了倒傳遞網路和學習向量量化網路。在倒傳遞網路方面，我們建一三層網路架構，輸入層含 1146 個結點(即關鍵詞數)，輸出層含 56 個結點(即類別數)，隱藏層含 50 個結點(測試結果為最佳)，以訓練資料訓練網路，經過 720,000 次(以一筆訓練資料為一次)，其實驗結果最佳：訓練資料召回率為 70.7%，測試資料為 57.1%。在學習向量量化網路方面，網路也為一三層之架構，輸入、輸出層同倒傳遞網路，隱藏層結點數取訓練資料的十分之一，共有 616 個結點(每一類別有 11 個聚類中心點)，訓練資料召回率為 86.9%，測試資料為 50.5%。

此外，我們並實驗了傳統的機率模式、向量模式、和 Hamill 的分類結果以為比較。在機率模式方面，訓練資料召回率為 82.7%，測試資料為 39.4%；在向量模式方面，訓練資料召回率為 55.4%，測試資料為 43.0%；在 Hamill 實驗中，訓練資料召回率為 63.6%，測試資料為 55.3%。我們將以上實驗結果表列如下：

表一：各種文件分類實驗方法的召回率

文件分類實驗方法	訓練資料	測試資料
倒傳遞網路	70.7%	57.1%
學習向量量化網路	86.9%	50.5%
機率模式	82.7%	39.4%
向量模式	55.4%	43.0%
Hamill 方法	63.5%	55.3%

四、結果與討論

在本實驗中，我們只採用了論文的題目和出處當作文件的素描，最主要是要和 Hamill 的實驗方法相比較，因為他的實驗只採用了論文題目。加入論文出處主要是這項資訊明顯地提升了分類的正確率，這可從下列數據得知：以 Hamill 方法為例，當我們的實驗資料不加入論文出處時，訓練資料召回率只有 56.3%，測試資料為 47.3%。(在 Hamill 的原本論文中[3]，以論文題目為文件素描時，測試資料僅有 45.0%的召回率。)當加入論文出處時，訓練資料召回率明顯提升為 63.5%，測試資料為 55.3%。因此，論文出處是項重要的分類特徵，它可大幅提升文件辨識正確率約達 10%。故本研究除了文件題目外，也將其出處當作文件的素描。

在前一節的實驗數據(表一)顯示，類神經網路在文件自動分類的表現均優於傳統的機率模式、向量模式、和 Hamill 方法，在訓練資料召回率方面，學習向量量化網路高達 86.9%；而在測試資料方面，倒傳遞網路可達 57.1%。此等顯示了類神經網路在文件自動分類的優越性。

在本實驗中，倒傳遞網路在訓練資料及測試資料的召回率均優於 Hamill 的方法。雖然如此，其訓練資料召回率只達到 70.7%，測試資料召回率只達到 57.1%，此可能為實驗用文件素描只包含論文題目

及其出處，平均每一篇文件只包含了 4.8 個關鍵詞，由於資訊相當有限，故在各種實驗中，測試資料召回率無法超越 60%。我們並分析了一下，以倒傳遞網路實驗為例，辨識正確及錯誤的文件中其所包含關鍵詞數，發現在訓練資料中辨識正確的文件其所含關鍵詞數約為 5.0 個，而錯誤文件關鍵詞數約為 4.1 個；在測試資料中辨識正確的文件其所含關鍵詞數約為 5.0 個，而錯誤文件關鍵詞數約為 4.4 個。由此觀察，文件的關鍵詞數越多，越有助於辨識之正確性。

另外，由於所蒐集的 6424 篇文件，並非平均分佈在 56 個類別中。有些類別所含文件高達三、四百篇，有些類別僅含一、二十篇。在倒傳遞網路的實驗中，類別中所含文件數少於三十篇者，幾乎無法正確辨識，這可能是因為類別的文件數相差太懸殊，以致於在訓練中，連結加權值的修正傾向於含文件較多之類別，以是之故，含較少文件數之類別辨識率相對就要減低許多。

五、計畫成果自評

在本計畫中，我們採用了論文的題目和其出處當作文件的素描。我們以類神經網路中之倒傳遞網路及學習向量量化網路作文件自動分類之實驗，其結果均較傳統之機率模式、向量模式、和 Hamill 方法為佳，已達成預期之目標。在原計畫案中，我們亦提出比較機率神經網路之實驗，然所使用之 NeuralWorks 軟體須將實驗資料作進一步正規化處理，因較複雜，故本實驗將之省略。雖然如此，本研究報告中加入了 Hamill 實驗方法以為比較，此為原計畫中所無。本研究顯示了類神經網路的強大學習能力，超過傳統的文件分類方法，此實驗結果應可發表於相關的學術會議[1]或期刊中。

[1] 本計畫報告將於「一九九八分散式系統技術及應用研討會」中發表，民國 87 年 5 月 14-15 日，國立成功大學。

六、參考文獻

[1] M.J. Blosseville, M.J. Hebrail, M.G. Monteil, and N. Penot, "Automatic Document Classification: Natural Language Processing, Statical Analysis, and Expert System Techniques Used Together," in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, June 21-24, 1992, pp. 51--58.

[2] H. Borko and M. Bernick, "Automatic Document Classification," *Journal of the ACM*, Vol. 10, No. 1, 1963, pp. 151--162.

[3] K.A. Hamill and A. Zamora, "The Use of Titles for Automatic Document Classification," *Journal of the American Society for Information Science*, Vol. 31, November 1980, pp. 396--402.

[4] P.S. Jacobs, "Joining Statistics with NLP for Text Categorization," in *Proceedings of the 3rd Conference on Applied Natural Language Processing*}, Trento, Italy, March 31-April 3, 1992, pp. 178--185.

[5] P.S. Jacobs, "Using Statistical Methods to Improve Knowledge-Based News Categorization," *IEEE Expert*, Vol. 8, No. 2, April 1993, pp. 13--23.

[6] G. Kar and L.J. White, "A Distance Measure for Automatic Document Classification by Sequential Analysis," *Information Processing and Management*, Vol. 14, 1978, pp. 57--69.

[7] T. Kohonen, "The Self-Organizing Map," *Proceedings of IEEE*, Vol. 78, No. 9, 1990, pp. 1481--1490.

[8] K.L. Kwok, "The Use of Title and Cited Titles as Document Representation for Automatic Classification," *Information and Management*, Vol. 11, 1975, pp. 201--206.

[9] K.J. MacLeod and W. Robertson, "A Neural Algorithm for Document Clustering," *Information Processing and Management*, Vol. 27, No. 4, 1991, pp. 337--346.

[10] M.E. Maron, "Automatic Indexing: An Experimental Inquiry," *Journal of the ACM*, Vol. 8, 1961, pp. 404--417.

[11] NeuralWare, Inc., *NeuralWorks Reference Guide*, Pittsburgh, NeuralWare, Inc., 1996.

[12] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning Internal Representation by Error Propagation," in *Parallel Distributed Processing*, Vol. 1, pp. 318--362.

[13] Y. Yang, "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ireland, 1994, pp. 13--22.